DOI: 10.1111/ddi.12668

BIODIVERSITY RESEARCH

WILEY Diversity and Distributions

In search of relevant predictors for marine species distribution modelling using the MarineSPEED benchmark dataset

Samuel Bosch[1,2] | Lennert Tyberghein[1] | Klaas Deneudt[1] | Francisco Hernandez[1] | Olivier De Clerck[2]

[1]Flanders Marine Institute (VLIZ), Ostend, Belgium

[2]Research Group Phycology, Biology Department, Ghent University, Ghent, Belgium

Correspondence
Samuel Bosch, Flanders Marine Institute (VLIZ), Ostend, Belgium.
Email: mail@samuelbosch.com

Funding information
Seventh Framework Programme, Grant/Award Number: SEAS-ERA/INVASIVES SD/ER/010; ERANET INVASIVES, Grant/Award Number: EU FP7 SEAS-ERA/INVASIVES SD/ER/010; Ghent University, the Hercules Foundation and the Flemish Government—department EWI

Editor: Alexandra Syphard

Abstract

Aim: Ideally, datasets for species distribution modelling (SDM) contain evenly sampled records covering the entire distribution of the species, confirmed absences and auxiliary ecophysiological data allowing informed decisions on relevant predictors. Unfortunately, these criteria are rarely met for marine organisms for which distributions are too often only scantly characterized and absences generally not recorded. Here, we investigate predictor relevance as a function of modelling algorithms and settings for a global dataset of marine species.

Location: Global marine.

Methods: We selected well-studied and identifiable species from all major marine taxonomic groups. Distribution records were compiled from public sources (e.g., OBIS, GBIF, Reef Life Survey) and linked to environmental data from Bio-ORACLE and MARSPEC. Using this dataset, predictor relevance was analysed under different variations of modelling algorithms, numbers of predictor variables, cross-validation strategies, sampling bias mitigation methods, evaluation methods and ranking methods. SDMs for all combinations of predictors from eight correlation groups were fitted and ranked, from which the top five predictors were selected as the most relevant.

Results: We collected two million distribution records from 514 species across 18 phyla. Mean sea surface temperature and calcite are, respectively, the most relevant and irrelevant predictors. A less clear pattern was derived from the other predictors. The biggest differences in predictor relevance were induced by varying the number of predictors, the modelling algorithm and the sample selection bias correction. The distribution data and associated environmental data are made available through the R package MARINESPEED and at http://marinespeed.org.

Main conclusions: While temperature is a relevant predictor of global marine species distributions, considerable variation in predictor relevance is linked to the SDM set-up. We promote the usage of a standardized benchmark dataset (MarineSPEED) for methodological SDM studies.

KEYWORDS

benchmark dataset, ecological niche modelling, marine, spatial cross-validation, species distribution modelling, variable importance

1 | INTRODUCTION

Climatological conditions are currently changing at an unprecedented rate and anthropogenic activities displace species out of their native area across the globe (Walther et al., 2009). Both processes have the potential to alter biological communities and reduce ecosystem services. Knowing under which environmental conditions species may maintain or establish viable populations therefore is more critical than ever. Species distributions are increasingly modelled for conservation and ecological purposes. A better understanding of mechanisms shaping species distributions allows for more accurate predictions of future distributions of species in a rapidly changing world (Franklin, 2009).

A mechanistic link between the abiotic factors and the species distributions is traditionally gleaned from physiological studies subjecting individuals to various environmental conditions and assessing their reaction norms. However, not all species lend themselves equally well to ex situ experiments. Also, the experimental set-up may only approximate realistic environmental conditions to a limited degree. Furthermore, physiological studies typically require prior knowledge on the ecological factors governing distribution ranges (Kearney & Porter, 2009). Given these difficulties, species distribution modelling (SDM), alternatively known as ecological niche modelling (ENM), offers an attractive alternative (Elith, Kearney, & Phillips, 2010). SDM correlates species occurrences, and optionally absences, with environmental data to create an estimation of the ecological niche and a projection in geographic space of this niche (Austin, 2002). The obvious advantage of correlative SDMs is that they require little knowledge of the mechanistic links between organisms and their environments. On the other hand, transferability of correlative models into novel areas or even for the same area in time is possibly compromised because of non-analogous climatic conditions. In such cases, experimental data on physiologically meaningful predictors present a significant added value (Fitzpatrick & Hargrove, 2009; Randin et al., 2006).

Thanks to the availability of an increasing number of online distribution records (e.g., OBIS, GBIF), pre-processed environmental data layers (e.g., WorldClim, Climond, Bio-ORACLE, MARSPEC) and modelling algorithms accessible through various statistical packages, SDM has become a widely applied technique in ecology and conservation biology (Pacifici et al., 2017). Studies on general SDM theory and methodology, however, focus mostly on terrestrial environments (reviewed in Elith & Leathwick, 2009; Franklin, 2009; Peterson et al., 2011). A minority of papers specifically address distribution modelling methods in the marine environment: presence-only algorithms (Beaugrand, Lenoir, Ibañez, & Manté, 2011; Cheung, Lam, & Pauly, 2008; Ready et al., 2010), algorithm comparisons (MacLeod, Mandleberg, Schweder, Bannon, & Pierce, 2008; Palialexis, Georgakarakos, Karakassis, Lika, & Valavanis, 2011; Šiaulys & Bučas, 2012), 3D modelling (Bentlage, Peterson, Barve, & Cartwright, 2013), rare species (Stirling, Boulcott, Scott, & Wright, 2016), joint SDMs (Torres, Read, & Halpin, 2008), ensemble modelling (Downie, von Numers, & Boström, 2013), scale effects (Nyström Sandman, Wikström, Blomqvist, Kautsky, & Isaeus, 2013; Pittman & Brown, 2011), null models (Merckx, Steyaert, Vanreusel, Vincx, & Vanaverbeke, 2011), model selection (Verbruggen et al., 2013), pseudo-absence generation (Coro et al., 2016; Huang, Brooke, & Li, 2011) and predictor datasets (Sbrocco & Barber, 2013; Tyberghein et al., 2012).

Although the importance of selecting biologically relevant predictors, and its impact on model uncertainty and transferability has been highlighted by several studies (Araújo & Guisan, 2006; Barry & Elith, 2006; Braunisch et al., 2013; Petitpierre, Broennimann, Kueffer, Daehler, & Guisan, 2017; Synes & Osborne, 2011; Verbruggen et al., 2013), to date no comprehensive study on the relevance of the predictors of marine species distributions across taxa has been performed. But, note that Bradie and Leung (2016), in their meta-analysis on variable importance from MaxEnt SDMs, included a limited set of marine species. These authors found that temperature and to a smaller extent bathymetry and salinity contributed most to marine species distribution models. While the impact of geographic scale, algorithm and pseudo-absence selection on the importance of predictors has been addressed to some degree (Bucklin et al., 2015; Elith et al., 2010; Nyström Sandman et al., 2013; VanDerWal, Shoo, Graham, & Williams, 2009), the impact of these and other aspects of SDM has not been studied on a global scale.

In this study, we created the Marine SPEcies with Environmental Data (MarineSPEED) dataset. This benchmark dataset, containing distribution records belonging to 514 well-studied taxa with a broad taxonomic, climatologic and geographic diversity, is used to investigate marine predictor relevance under an array of modelling parameters and algorithms. With this, we aim to answer two questions: (1) what are the most relevant predictors of marine species distributions and (2) which parts of the SDM process impact the relevance of predictors the most. Additionally, this study aims to promote the usage of benchmark datasets in methodological SDM studies as this allows for reproducible and comparable results.

2 | METHODS

2.1 | Species data

For the marine species benchmark dataset, we selected species from an array of taxonomic groups, climatological preferences and distribution patterns. We aimed to include species that are well studied in terms of their distribution and that often would classify as iconic species. For a species to be considered, we required the availability of at least 100 distribution records.

Species distribution records were collected from the Ocean Biogeographic Information System (OBIS; http://iobis.org, accessed February 2016), from the Global Biodiversity Information Facility (GBIF; http://gbif.org, accessed January 2016), the Reef Life Survey (RLS; http://reeflifesurvey.com, accessed February 2016) and for a few species via personal communications. For downloading the records from OBIS and GBIF, the R (R Core Team, 2016) clients ROBIS (Provoost, Bosch, & Appeltans, 2016) and RGBIF (Chamberlain, Boettiger, Karthik, Barve, & Mcglinn, 2016) were used, respectively. A list of data sources is found in Appendix S1. The distribution records were subsequently filtered until only one record remained in each cell

of an equal-area grid with a per cell area of 25 square kilometres. This step eliminates duplicated records from different data sources and limits the number of records from repeated sampling events in the same area. We also removed records located within the land mask of the environmental data. Finally, the distributions for all species were visually inspected and cross-checked with available distribution information to eliminate erroneous records.

We collected for each species taxonomic and functional group information from the World Register of Marine Species (WoRMS Editorial Board, 2016). The "functional group" trait divides species into three groups reflecting their habitat: benthos, nekton and plankton (zooplankton and phytoplankton). For species lacking trait data in WoRMS, this information was derived from FishBase (Froese & Pauly, 2017) and SeaLifeBase (Palomares & Pauly, 2017) whereby all seafloor-associated species were classified as benthos (i.e., sessile, reef-associated or demersal species), other free-swimming species as nekton and drifting species as plankton. In addition, species were categorized as oceanic if more than five per cent of their records are located outside the marine ecoregions. Else, species were considered as neritic. Last, we classified organisms according to latitudinal zones ("polar," "temperate," "tropical"). Thereto, we checked for the presence of at least five per cent of all occurrence records of a species in each latitudinal zone of the marine ecoregions classification by Spalding et al. (2007).

2.2 | Environmental data

The distribution records in the MarineSPEED dataset were linked to 68 monthly and annual environmental variables for the current climate available from Bio-ORACLE (Tyberghein et al., 2012) and MARSPEC (Sbrocco & Barber, 2013) with a spatial resolution of 5 arcmin using the R package SDMPREDICTORS (Bosch, Tyberghein, & De Clerck, 2016). These environmental data include variations of sea surface temperature, salinity, bathymetry, nutrients and other predictors of marine species distributions.

2.3 | Background data

Most presence-only SDM methods use background or pseudo-absence points for building models (Franklin, 2009). To facilitate the reproducibility of different studies using MarineSPEED, we included a set of 20,000 randomly sampled background points in the benchmark dataset. We also created a second set of target-group background points by randomly sampling 20,000 points from the full set of distribution records. The latter show the same bias as the occurrence records and therefore can be used to mitigate the effect of sample selection bias on presence-only species distribution models (Kramer-Schadt et al., 2013; Phillips et al., 2009; Syfert, Smith, & Coomes, 2013).

2.4 | Cross-validation splits

Cross-validation (CV) is a widespread strategy used to perform model selection while avoiding under- and overfitting models (Arlot & Celisse, 2010). We prepared CV folds for the species and background data using three different strategies. As a first strategy, we partitioned the data randomly in five folds (random CV). This strategy is easy to perform but has as disadvantage that it commonly results in an overestimated performance of the model because training and validation points selected from nearby locations will be dependent due to the effect of spatial autocorrelation (Bahn & McGill, 2007; Hijmans, 2012; Roberts et al., 2016). As CV only avoids overfitting when training samples are independent from the validation samples, this generally leads to the selection of complex models with poor transferability (Arlot & Celisse, 2010; Petitpierre et al., 2017; Verbruggen et al., 2013). The second (disc-based CV) and third (grid-based CV) splitting strategies take into account the spatial nature of the data. The fivefold disc-based strategy randomly samples a starting point and subsequently selects the nearest one-fifth of all distribution records to get the first fold. Then, the distribution record farthest away from the starting point is used as a new starting point and the nearest one-fifth of the distribution records are included to create the second fold. This process is repeated five times until all records are assigned to a fold. For the fourfold grid-based strategy, records are split into two sets based on their longitude using a random meridian as a dividing line. Then, these two halves are separately split in two equal parts using parallels. Additionally, ninefold grid-based sets were created using two meridians and parallels for splitting instead of one. By combining the disc- or grid-based CV strategies with the pairwise distance sampling method proposed by Hijmans (2012) to select the pseudo-absence points for the test set spatial sorting bias was eliminated and thus the effect of spatial autocorrelation on the performance evaluation suppressed (Bahn & McGill, 2007; Roberts et al., 2016). To remove false negatives in the training sets of the spatial cross-validation sets, we excluded background points from the training sets that are within 200 km of test occurrences.

2.5 | Predictor relevance

To find out which predictors are most relevant for the set of species in MarineSPEED, we ranked distribution models fitted for all combinations of predictors from multiple correlation groups. In addition, we added variation at the different steps of the model creation to assess the variability in predictor relevance under different model set-ups (Figure 1).

Following the methodology from Barbet-Massin and Jetz (2014), who identified relevant predictors of bird distributions, distributions were modelled for all combinations of three, four and seven environmental predictors selected from eight correlation groups. After filtering the initial set of 68 predictors down to 19 based on a Pearson's product moment correlation coefficient larger than 0.95, we created correlation groups with the R package SDMPREDICTORS by grouping all predictors for which some or all of the predictors have an absolute Pearson's product moment correlation coefficient larger than 0.7 (Barbet-Massin & Jetz, 2014; Dormann et al., 2013). This resulted in eight correlation groups of which six predictors form a group on their own (shore distance, bathymetry, SST (range), calcite,

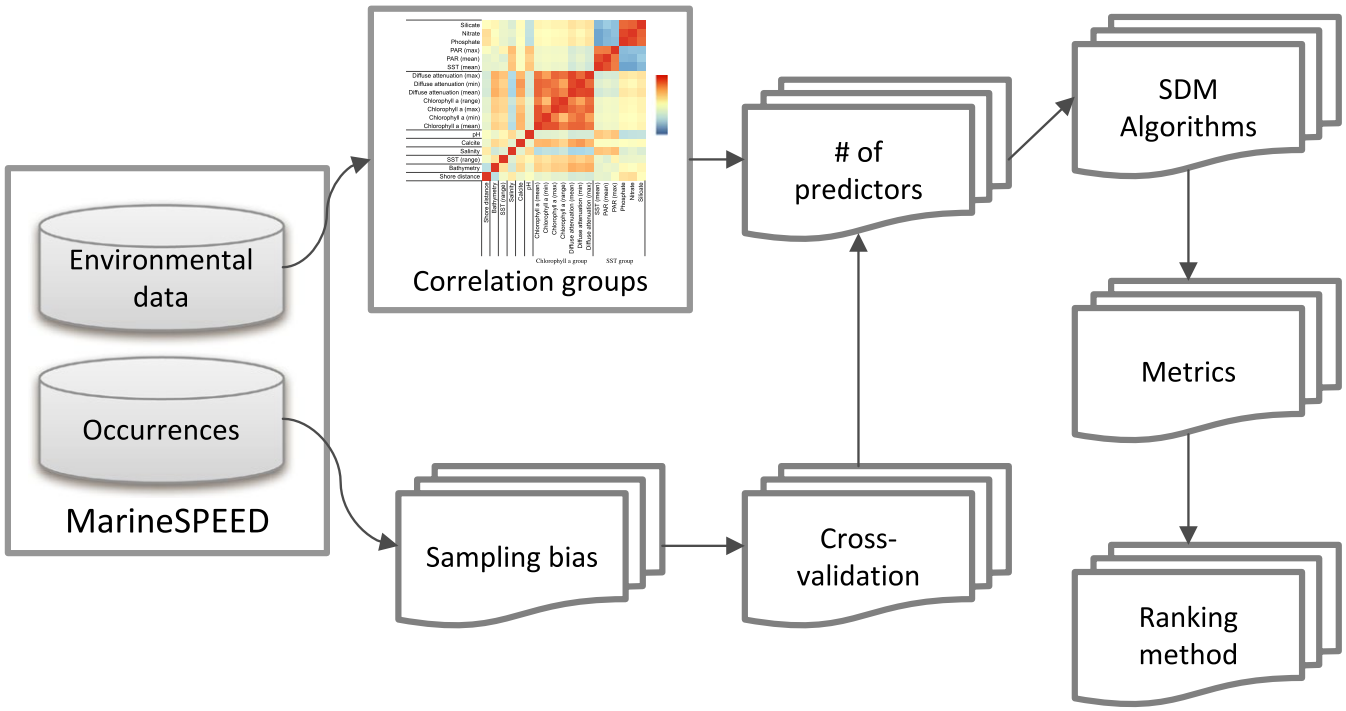

FIGURE 1 Overview of the predictor selection analysis. Starting from 19 environmental predictors, from Bio-ORACLE and MARSPEC, correlation groups where created. For these groups, all possible combinations of models with three, four and seven predictors were generated. After optional sample selection bias mitigation, occurrence records and background points were split in random or spatial cross-validation folds. SDMs were build using four algorithms (random forests, MaxEnt, generalized linear models and Bioclim) and evaluated using the area under the curve of the receiver operating characteristic (AUC) and the point-biserial correlation (COR). Predictors were ranked based on the performance of the models they were included in

salinity, pH), seven predictors belong to the "Chlorophyll *a* group," grouping chlorophyll *a* and diffuse attenuation (mean, minimum, maximum and/or range) related variables. The last six predictors form the "SST group" with variations of sea surface temperature (SST), photosynthetically active radiation (PAR), phosphate, nitrate and silicate. For a full overview of the different environmental predictors used and the correlation group they belong to, we refer to Figure 2 and to Table S1 in Appendix S3. Additionally, the performance of the mean SST was compared with the performance of minimum and maximum SST.

SDMs were fitted using four commonly used algorithms: Bioclim (Booth, Nix, Busby, & Hutchinson, 2014), Generalized Linear Model (GLM), Maximum Entropy modelling (Maxent, Phillips, Dudík, & Schapire, 2004) and Random Forests (RF, Breiman, 2001). We used the dismo (Hijmans, Phillips, Leathwick, & Elith, 2016) and randomForest (Liaw & Wiener, 2002) packages in R for fitting Bioclim and MaxEnt and random forest models, respectively. For all algorithms, the default settings were used and GLMs were run with only linear features. To evaluate potential differences in model performance due to selection of mean vs. minimum or maximum temperature, we repeated the analyses allowing for seven predictors using all four algorithms and a disc-based cross-validation.

Three variations of sample selection bias correction were performed: (1) no correction, (2) spatial thinning (50 km) with the R package SPTHIN (Aiello-Lammens, Boria, Radosavljevic, Vilela, & Anderson, 2015) and a target-group background (Phillips et al., 2009). Performance of the models was evaluated using random as well as

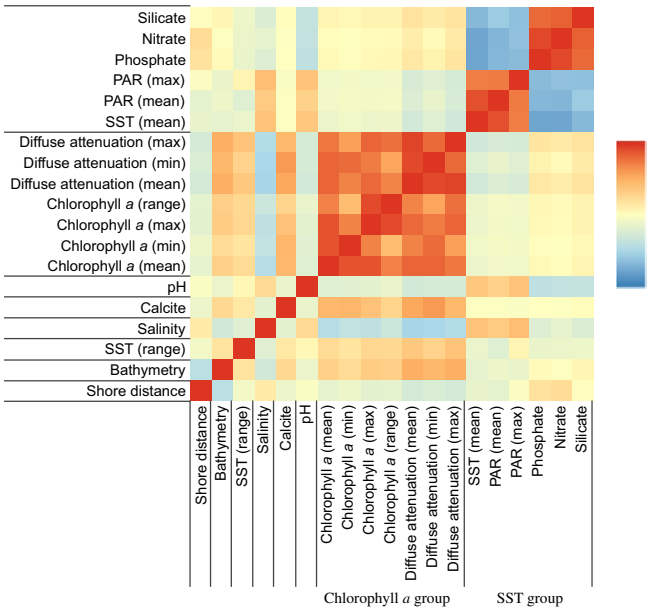

FIGURE 2 Correlation matrix for all environmental predictors considered for the predictor selection analysis, grouped by correlation group. Note that for creating the correlation groups, predictors are grouped when the absolute correlations between two or more members of a correlation group are higher than 0.70

spatial disc-based cross-validation. In total, six million models were fitted and evaluated using the area under the receiver operating characteristics (ROC) curve (AUC) (Hanley & McNeil, 1982), and the

point-biserial correlation (COR) (Elith et al., 2006; Zheng & Agresti, 2000) on the UGent High Performance Cluster.

Per species, the modelling options described above resulted in a list of AUC or COR values. The mean and median AUC and COR values for the models in which a specific predictor was used were calculated and ranked across predictors. In addition, we used rank centrality (Negahban, Oh, & Shah, 2017), an iterative algorithm for rank aggregation based on pairwise–wise comparisons of the performance of all models in which the different predictors were used. Rank centrality produces a score for each predictor which is then ranked to obtain the final predictor rankings for each model set-up, evaluation metric and species combination. The predictor relevance was determined by calculating the percentage of species for which the predictor ranked in the top five for each modelling option, evaluation metric and ranking method.

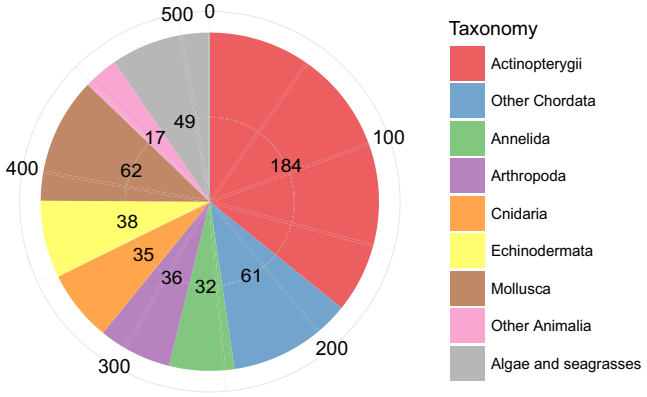

FIGURE 3 Taxonomic composition of the MarineSPEED dataset on level kingdom, phylum or class. For the kingdom Animalia, the most abundant phylum Chordata was split up into the Actinopterygii and other Chordata, the kingdom Plantae was left as one whole and labelled as algae and seagrasses. Numbers represent the number of species in each taxonomic group

3 | RESULTS

3.1 | Benchmark dataset

The MarineSPEED benchmark dataset is composed of 514 species with an original two million distribution records which have been filtered down on a 25-km^2 grid to nearly nine hundred thousand records. On a species level, the median number of filtered distribution records is 506 with a minimum of 52 and a maximum of 45,469. A summary of the taxonomic and biogeographic information per species is available in Appendix S2.

A total of 18 different phyla are included in MarineSPEED (Figure 3), with as best represented phyla: Chordata (245 species), Mollusca (62 species), Echinodermata (38 species), Arthropoda (36 species) and Annelida (32 species). The phylum Chordata is mostly represented by the class Actinopterygii (184 species), and to a lesser extent Elasmobranchii (20 species) and Mammalia (18 species). Marine primary producers, various groups of algae and seagrasses, are represented by 49 species from five phyla. When classifying species into functional groups, 395 species are associated with the seafloor (benthos), while 87 species are free swimming (nekton) and 32 species are planktonic. While we aimed to select species from different parts of the world, a bias towards well-researched areas (e.g., the North-Atlantic and Australia) was unavoidable (Figure 4). Likewise, coastal areas are overrepresented compared to open ocean habitats. On a latitudinal scale, temperate regions are the most represented with 173 species. Ninety-one species only occur in the tropics and 11 species are restricted to polar regions; 72 species have more than five per cent of their records in the open ocean.

The predefined spatial cross-validation splits all increase the distance between test points and their nearest training point as

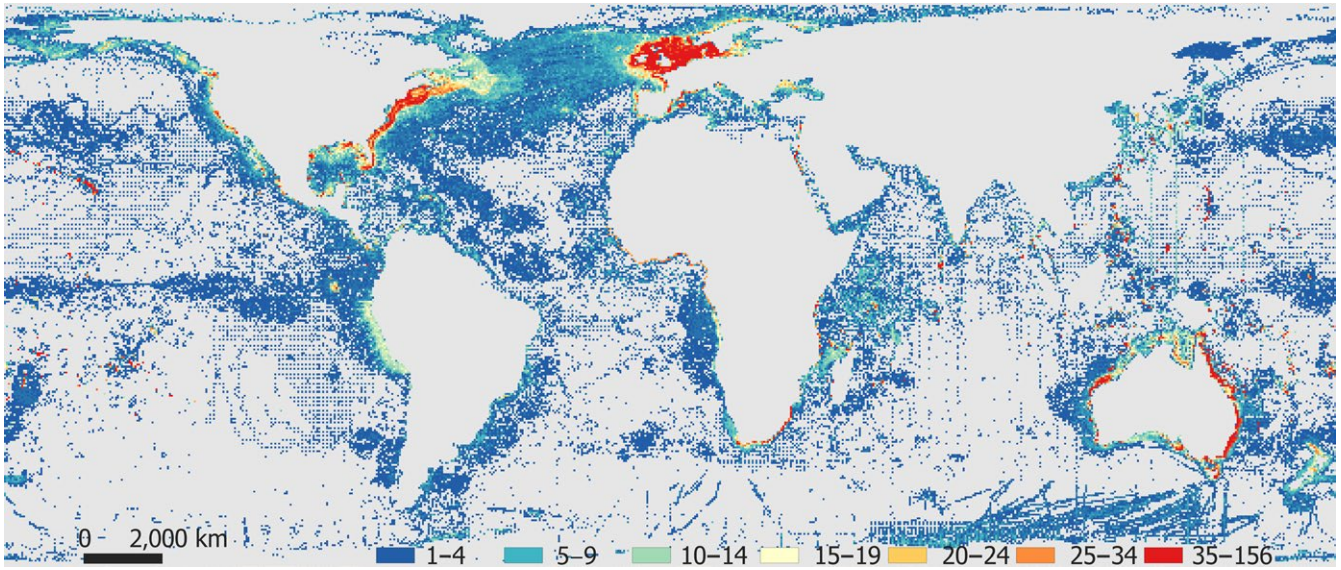

FIGURE 4 Map of the number of species occurring in each cell of an equal-area grid with a per cell area of 25 km^2 (Behrmann cylindrical equal-area projection)

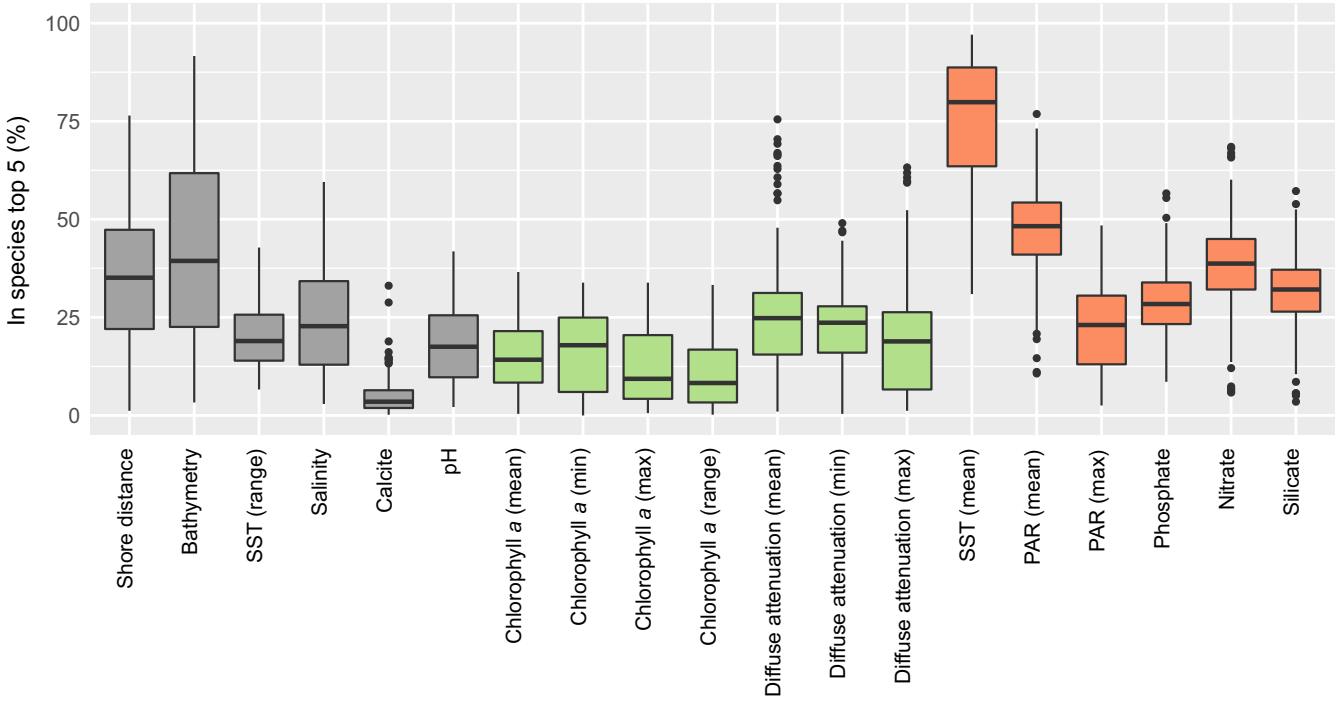

FIGURE 5 Percentage of species a predictor has a top 5 ranking in the different model set-ups. In grey are the predictors that form a correlation group on their own, in green the predictors from the "Chlorophyll *a* group" and in red the predictors from the "SST group." The results are aggregated from all possible variations. For a detailed view on the different dimensions of the variations, we refer to Tables 1–3, and to the following plots in Appendix S3: modelling algorithms (Fig. S4), evaluation metrics (Fig. S5), ranking methods (Fig. S6), cross-validation strategies (Fig. S7), predictor counts (Fig. S8), sampling bias mitigation methods (Fig. S9), cross-validation folds (Fig. S10), taxonomic groups (Fig. S11) functional groups (Fig. S12), zones (Fig. S13) and ecoregions (Fig. S14)

compared to random splits (Fig. S1 in Appendix S3). Examples of the various cross-validation strategies are visualized for *Didemnum maculosum* Milne Edwards and *Polycarpa aurata* Quoy & Gaimard in Figs S2 and S3, respectively, in Appendix S3.

3.2 | Predictor relevance

A first set of analyses exploring the selection of relevant predictors (Figure 5) highlights the importance of mean sea surface temperature, SST (mean), as the most relevant predictor of species distributions in the MarineSPEED benchmark dataset. This result appears robust regardless of modelling algorithm, sample selection bias correction, cross-validation, number of predictors, evaluation metric, ranking method and taxonomic groups. Analyses whereby SST (mean) was replaced by either SST (max) or SST (min) did not alter the importance of SST as predictor in the models. Neither did these changes affect the performance of the models as demonstrated by AUC or COR values which were virtually identical (Fig. S15). At the other end of the spectrum, calcite is apparently irrelevant as a predictor for most of the species distributions. As for the other predictors, however, there is substantial variation across species and modelling parameters.

Among the different algorithms, GLMs with linear features caused the most variation in the predictor top 5 rankings with a particularly strong effect on SST (mean) with a minimal decrease of 28% in the median percentage of species with SST (mean) in the top 5 ranking (Table 1). Conversely, in GLMs bathymetry was selected at least 26% more. The difference between the two evaluation metrics AUC and COR on the other hand was fairly limited with salinity displaying the largest difference. Finally, the ranking method showed very small differences between the mean and median ranking algorithm. The rank centrality algorithm consistently ranked the predictors from the "Chlorophyll *a* group" as less relevant, while increasing the ranking of salinity (+16%), bathymetry (+15%), pH (+13%) and shore distance (+13%).

When comparing the results of CV splitting strategies, number of predictors, sampling bias mitigation and fold number (Table 2), we can conclude that the number of predictors allowed in the model has the largest effect. Increasing the number of allowed predictors from 3 to 7 causes a decline in the relevance of bathymetry (−31%) and shore distance (−26%) while increasing the relevance of PAR (max) (+17%), diffuse attenuation (max) (+14%) and chlorophyll *a* (max and range) (+13%). The second largest effect is caused using a target-group background to mitigate the effect of sampling bias on SDMs with a decrease of 25% for bathymetry and 15% for shore distance and an increase of 12% for nitrate. When using the disc-based CV strategy, the relevance of SST (mean) and salinity decreased with 19% and 10%, respectively. Using the second fold instead of the first fold, which was only performed for the random CV strategy, only yielded small differences in the top 5 predictors of the species.

While the relevance of most predictors is similar across taxonomic groups, some predictors exhibit large differences (Table 3). This is

TABLE 1 Median percentage of species for which a predictor has a top 5 ranking for the different set-up variations that have been calculated for all models

			Algorithm				Metric		Ranking method		
Group	Predictor	**All**	Bioclim	GLM	MaxEnt	RF	AUC	COR	Centrality	Mean	Median
	Shore distance	**35**	29	22	39	40	36	34	44	31	27
	Bathymetry	**39**	45	71	36	19	40	37	52	37	33
	SST (range)	**19**	14	24	19	18	18	19	26	16	16
	Salinity	**23**	16	15	25	37	18	26	33	17	16
	Calcite	**4**	4	5	3	3	3	4	6	2	3
	pH	**18**	8	24	14	23	17	18	26	12	13
Chlorophyll *a* group	Chlorophyll *a* (mean)	**14**	18	8	14	17	15	13	9	16	18
	Chlorophyll *a* (min)	**18**	22	4	21	21	17	18	6	22	22
	Chlorophyll *a* (max)	**9**	15	6	11	15	10	9	5	17	19
	Chlorophyll *a* (range)	**8**	11	7	9	13	8	9	3	13	15
	Diffuse attenuation (mean)	**25**	21	44	24	24	23	26	10	27	27
	Diffuse attenuation (min)	**24**	22	30	22	21	22	23	9	25	25
	Diffuse attenuation (max)	**19**	12	37	10	16	18	19	7	23	23
SST group	SST (mean)	**80**	79	51	89	86	79	78	79	79	78
	PAR (mean)	**48**	53	49	48	41	46	49	51	46	46
	PAR (max)	**23**	22	30	20	15	20	24	26	17	22
	Phosphate	**28**	32	23	27	32	29	27	33	26	26
	Nitrate	**39**	41	31	41	44	41	33	41	38	37
	Silicate	**32**	27	29	32	36	32	31	36	29	31

First column shows the results for all models. The next four columns show the results for the different modelling algorithms: climate envelope model (Bioclim), Generalized Linear Model (GLM), Maximum Entropy modelling (MaxEnt) and Random Forests (RF). Followed by two columns showing the breakdown for the evaluation metrics used: area under the receiver operating characteristics curve (AUC) and the point-biserial correlation (COR). The last three columns show the results for the ranking methods: ranking using the Rank Centrality algorithm or ranking of the mean or median performance of predictors. Green indicates a low percentage and thus a small relevance, yellow indicates a medium relevance and red indicates a high relevance.

especially the case for shore distance, bathymetry and SST (range) with differences between the minimum and maximum of 55%, 40% and 33%, respectively. Despite these overall patterns in the median ranking values, we see that the spread of the predictor relevance within taxonomic groups is large (Fig. S11).

Table 4 presents the results related to species traits: functional group, neritic vs. oceanic zone and ecoregion. Some clear trends are visible whereby shore distance, bathymetry and to a lesser extent PAR (mean) are comparatively more relevant predictors for benthic species distributions, less relevant for nekton and least relevant for plankton. For mean and minimum diffuse attenuation, we notice an inverse trend with a higher relevance for plankton in comparison with nekton and benthos. With respect to the zone trait, we see that shore distance (−21%) and bathymetry (−14%) are less relevant for oceanic species, while phosphate (+15%), nitrate (+13%) and silicate (+15%) are more relevant. The results from the latitudinal zone trait show clear differences in predictor relevance for multiple predictors. For some predictors such as SST (range), nitrate and phosphate, the relevance for temperate species clearly deviates from that for polar and tropical species. For boxplots of the relevance of the predictors for the different variations in model set-up, taxonomic groups and species traits we refer to Figs S4 to S14 in Appendix S3.

3.3 | Data access

While distribution maps for all species can be consulted and all data are downloadable in an R Shiny interface (Chang, Cheng, Allaire, Xie, & McPherson, 2016) at <http://marinespeed.org>, we opted to also create the MARINESPEED R package allowing for easy usage of the data (Table 5). The first step, after installation from CRAN and loading the library, is to run the function "list_species" which returns the scientific names and WoRMS identifiers for all species. Additional information on the taxonomy and latitudinal zones can be viewed using the "species_info" function. To run a function for all species, either the "lapply_species" or the "lapply_species_kfold" function can be used. Alternatively, if you only need data for specific species, the "get_occurrences" and "get_fold_data" methods can be used. Lower level functions for loading background data and creating cross-validation splits are also available.

4 | DISCUSSION

Species distribution modelling is widely used to identify areas that are ecologically suitable for the presence of species under past,

TABLE 2 Overview of the median percentage of species for which a predictor has a top 5 ranking for the different set-up variations that have been calculated for a subset of the models

			CV splitting strategy		Predictor count			Sampling bias mitigation			Fold number	
Group	Predictor	**All**	Disc	Random	3	4	7	None	spThin	Target-group	1	2
	Shore distance	**35**	35	30	56	55	30	30	27	12	30	35
	Bathymetry	**39**	42	34	65	62	34	34	33	8	34	37
	SST (range)	**19**	15	21	19	24	21	21	18	18	21	11
	Salinity	**23**	13	23	22	28	23	23	23	28	23	20
	Calcite	**4**	9	3	3	3	3	3	3	2	3	3
	pH	**18**	11	17	17	17	17	17	16	27	17	16
Chlorophyll *a* group	Chlorophyll *a* (mean)	**14**	15	18	12	12	18	18	15	22	18	19
	Chlorophyll *a* (min)	**18**	21	17	18	15	17	17	19	16	17	17
	Chlorophyll *a* (max)	**9**	16	16	3	4	16	16	17	19	16	14
	Chlorophyll *a* (range)	**8**	17	15	2	4	15	15	15	14	15	12
	Diffuse attenuation (mean)	**25**	18	26	24	24	26	26	26	28	26	27
	Diffuse attenuation (min)	**24**	24	24	25	21	24	24	25	19	24	24
	Diffuse attenuation (max)	**19**	18	20	6	8	20	20	21	25	20	22
SST group	SST (mean)	**80**	59	78	85	84	78	78	80	85	78	76
	PAR (mean)	**48**	46	50	37	47	50	50	51	59	50	49
	PAR (max)	**23**	34	25	8	12	25	25	25	25	25	23
	Phosphate	**28**	32	26	28	27	26	26	27	28	26	30
	Nitrate	**39**	36	33	42	38	33	33	34	46	33	44
	Silicate	**32**	36	35	29	29	35	35	32	29	35	29

In this table, only results from set-ups that have been done for both options are shown. First column shows the results for all models, the next two columns show the results for the fivefold random and disc-based spatial cross-validation splitting strategies, and the next three columns show the breakdown for the number of predictors used in the models. The next three columns show the impact of using sampling bias mitigation techniques on the predictor relevance by comparing doing nothing with performing spatial thinning (spThin) and with using a background from a sample of all species records (target-group background). The last two columns show the results for the first and the second fold of a fivefold random cross-validation. Green indicates a low percentage and thus a small relevance, yellow indicates a medium relevance and red indicates a high relevance.

current and future climates. Most studies concentrate, however, on terrestrial environments, while marine species distribution modelling kicked off comparatively late (Robinson et al., 2011). A direct consequence of the relative scarcity of marine SDM studies is that most of the methodological progress in SDM is biased towards terrestrial studies, despite marine environments being significantly different with respect to the ecological factors that control distributions and their spatio-temporal variation. These differences raise questions with respect to the environmental predictor relevance and the effects of model algorithms and settings on predictor relevance. By fitting presence-only SDMs for all combinations of predictors from different correlation groups, we assessed the predictor relevance and the variation therein for marine species distributions. To this end, we created a benchmark dataset (MarineSPEED) which bundles marine species distributions of 514 taxa and associated environmental variables.

4.1 | Relevant predictors

SST (mean) is the most relevant predictor of global marine species distributions, regardless of model algorithms and parameter settings. Our results corroborate the analyses by Belanger et al. (2012) who identified mean sea surface temperature as the most important single environmental predictor of biogeographic structure of marine benthic faunas. SST was also the only statistically significant predictor of species richness across species groups in the marine environment by Tittensor et al. (2010). These combined results support the idea that adaptation to thermal windows shapes both the distribution and diversity patterns of marine biota. The only groups that seem to defy this pattern are endothermic marine mammals that are able to decouple metabolic rates from ambient temperatures (Pörtner, 2002; Tittensor et al., 2010). The strong correlation of marine ectotherm distributions with temperature supports an underlying metabolic explanation to define the thermal tolerance range required for maintenance of a population of ectotherms in their natural environment. Work by Pörtner (2002) highlights the importance of specific upper or lower temperatures which mark a decrease in growth. Outside these temperatures, tolerance exists but becomes increasingly time-limited.

Given the importance of temperature thresholds, we investigated up to which extend long-term mean temperatures are able to predict distributions better than minima or maxima. In line with the high degree

TABLE 3 Median percentage of species for which a predictor has a top 5 ranking for the different set-up variations that have been calculated for all models and for some taxonomic groups

			Chordata	Other Animalia					Plantae
Group	Predictor	All	Actinopterygii	Annelida	Arthropoda	Cnidaria	Echinodermata	Mollusca	Algae and seagrasses
	Shore distance	**35**	42	16	11	66	32	29	44
	Bathymetry	**39**	49	42	33	54	51	31	14
	SST (range)	**19**	14	42	36	9	13	19	14
	Salinity	**23**	18	16	19	11	21	25	31
	Calcite	**4**	2	3	6	3	8	3	4
	pH	**18**	19	6	11	11	13	21	18
Chlorophyll *a* group	Chlorophyll *a* (mean)	**14**	11	9	17	9	16	15	18
	Chlorophyll *a* (min)	**18**	15	16	19	9	16	15	20
	Chlorophyll *a* (max)	**9**	9	5	11	6	13	10	10
	Chlorophyll *a* (range)	**8**	8	3	8	6	11	10	8
	Diffuse attenuation (mean)	**25**	17	31	33	11	18	27	35
	Diffuse attenuation (min)	**24**	17	31	25	9	21	23	29
	Diffuse attenuation (max)	**19**	19	9	17	14	21	18	18
SST group	SST (mean)	**80**	81	81	72	83	71	69	76
	PAR (mean)	**48**	52	41	42	57	45	47	33
	PAR (max)	**23**	18	28	33	9	21	24	18
	Phosphate	**28**	29	19	33	29	26	26	23
	Nitrate	**39**	43	22	36	47	34	35	24
	Silicate	**32**	25	41	36	14	29	35	39

Within the class Chordata and within the kingdom Animalia, taxa with few species were left out of this comparison. Green indicates a low percentage and thus a small relevance, yellow indicates a medium relevance and red indicates a high relevance.

of correlation between minimum, mean and maximum SST, the AUC or COR values of the models are virtually identical. Likewise, predictor relevance is also not affected. These conclusions result from broad-scale comparisons, which does not necessarily imply that the resulting models are completely identical. For example, it would be interesting to see whether minima or maxima or able to predict range edges more accurately than long-term mean SST values. In particular in geographic regions where minimum, mean and maximum SST are somewhat less correlated we would expect to see differences in prediction.

While bathymetry and shore distance are on average very relevant, there is considerable variance in the results, which might be because they are distal environmental predictors (Austin, 2002). In contrast to previous results (Bradie & Leung, 2016; Nyström Sandman et al., 2013), bathymetry was not the most important predictor, which can be explained by the global scale of our study. The importance of bathymetry has been shown to decrease with increased geographic scale (Nyström Sandman et al., 2013). Moreover, the relevance of bathymetry is strongly linked to the species taxonomy (see Tables 3 and 4 and Figs S11–S14). At the other end of the spectrum, calcite is rarely selected as a meaningful predictor. The irrelevance of calcite is consistent with the fact that only one study in the meta-analysis by Bradie and Leung (2016) used calcite as a predictor. The remaining predictors are on average less often included in the best scoring models, reflecting an overall reduced relevance towards predicting species distributions.

Despite this general trend, the variance in predictor relevance is relatively high across model algorithms and settings. The high variance when using different modelling algorithms is consistent with the results by Bucklin et al. (2015) who also demonstrated a significant interaction between predictor set and modelling algorithm. In particular, predictor selection under GLM deviates from the other algorithms. GLM-based models do not capture the relevance of SST (mean) very well. The lower relevance of SST in GLM models indicates that the global distribution of marine species is inadequately modelled by a linear relationship. Potentially, this effect can be mitigated by including polynomial features, an option which was not explored in the current analyses. In MaxEnt, with automatic selection of feature complexity and therefore yielding complex models, the relevance of SST (mean) is consistently high and displays hardly any variation. We expect that decreasing the complexity of the features fitted by MaxEnt will result in models more similar to GLM-based models. As for the other three algorithms, predictor selection seems to be largely consistent, echoing results of Barbet-Massin and Jetz (2014).

We also compared the predictor relevance under two different evaluation measures, AUC and COR, respectively. Although AUC, as an absolute measure for model performance, has been criticized earlier (Lobo, Jiménez-Valverde, & Hortal, 2010), its use is warranted here as we only compared relative AUC values and only modelled in a fixed geographic extent. Both AUC, which measures the ability to discern presences from background data, and COR, which provides a measure for the calibration of the model, showed very similar predictor rankings. This similarity is indicative for the generalizability of the results across model evaluation metrics.

Likewise, for most predictors the ranking method used and did not affect the predictor relevance. The rank centrality method consistently gave a lower ranking to all predictors from the "Chlorophyll *a* group." Although the Rank Centrality outperforms other popular ranking algorithms, ranking from pairwise comparisons is an active research fields

TABLE 4 Median percentage of species for which a predictor has a top 5 ranking for the different set-up variations that have been calculated for all models and traits

			Functional group			Zone		Ecoregion		
Group	Predictor	All	Benthos	Nekton	Plankton	Neritic	Oceanic	Polar	Temperate	Tropical
	Shore distance	**35**	39	24	13	38	17	13	25	49
	Bathymetry	**39**	44	26	13	40	26	39	25	60
	SST (range)	**19**	17	22	28	19	19	13	28	5
	Salinity	**23**	20	24	22	22	18	26	27	14
	Calcite	**4**	3	2	3	3	1	0	3	2
	pH	**18**	17	17	6	19	7	4	18	14
Chlorophyll *a* group	Chlorophyll *a* (mean)	**14**	12	16	19	13	17	17	17	10
	Chlorophyll *a* (min)	**18**	16	19	17	17	17	13	20	11
	Chlorophyll *a* (max)	**9**	8	11	9	9	10	4	10	8
	Chlorophyll *a* (range)	**8**	8	8	9	8	10	4	8	8
	Diffuse attenuation (mean)	**25**	22	30	44	24	25	30	33	12
	Diffuse attenuation (min)	**24**	21	27	34	23	24	22	31	10
	Diffuse attenuation (max)	**19**	18	16	16	19	15	13	16	20
SST group	SST (mean)	**80**	79	77	78	79	77	74	74	86
	PAR (mean)	**48**	49	45	34	48	44	26	42	56
	PAR (max)	**23**	21	29	19	22	19	17	25	14
	Phosphate	**28**	27	25	34	25	40	48	21	34
	Nitrate	**39**	39	33	31	36	49	57	27	49
	Silicate	**32**	28	39	41	28	43	52	38	16

For the functional group trait, benthos includes all seafloor-associated species, including demersal and reef-associated species; nekton includes all actively swimming pelagic species and plankton are all species unable to swim against a current. The neritic and oceanic zones were defined based on the ecoregion classification by Spalding (2007) whereby species having 5% or more of their distribution records outside of ecoregions are classified as oceanic. Species are a member of an ecoregion when at least 5% of its distribution records are situated in a polar, temperate or tropical ecoregion. Green indicates a low percentage and thus a small relevance, yellow indicates a medium relevance and red indicates a high relevance.

TABLE 5 Overview of the most important functions in the MARINESPEED R package

Function	Description
list_species	Get the list of scientific names and WoRMS identifiers for all species
species_info	Additional species information
lapply_species	Execute a function for all distribution records for multiple species
lapply_kfold_species	Execute a function for one or more pre-made CV folds for multiple species

Lower level functions for accessing occurrences, background data and creating CV folds are also available.

(Negahban et al., 2017). For instance Bradie and Leung (2016) used Microsoft's TrueSkill method, a Bayesian skill ranking system that generalizes the ELO chess ranking system (Herbrich, Minka, & Graepel, 2006), and other pairwise ranking methods have been recently proposed such as spectral ranking (Fogel, D'Aspremont, & Vojnovic, 2016) and sync rank (Cucuringu, 2016). A future study comparing these different ranking methods could lead to additional insights on the impact of the ranking algorithm on the predictor relevance.

The impact of cross-validation strategies was assessed using spatial disc-based and random sampling of training and testing sets. Using a spatial instead of a random data splitting strategy resulted in a lower relevance of SST (mean). This can be attributed to two different factors: (1) extrapolation and (2) scale effects. Firstly, spatial data splits frequently result in part of the SST range of a species not being included in the model, causing extrapolation artefacts during model validation (Roberts et al., 2016). While SST is in general the most relevant predictor, spatial validation may therefore lead to low evaluation scores and a lower relevance. In the marine environment, differences in surface temperature tend to be noticeable at comparatively large distances. Therefore, short distances between test presences and pseudo-absences will decrease the relevance of temperature as a predictor variable. The scale effect results from the average distance which tends to be smaller in spatial compared to random cross-validation. These results confirm that SST is especially relevant on a global scale but less so on a smaller scale (Nyström Sandman et al., 2013).

Restricting the number of predictors included in a model directly influences the relevance of the predictors. For most marine species, the relevance of bathymetry and shore distance diminishes when more predictors are included in the model. These predictors are only distally related to the suitability of an environment for species distributions, and therefore, the potential choice of more proximate predictors will result in their lower relevance in predictor-rich models. Inversely, predictors from the "Chlorophyll *a* group" are selected more, suggesting that if combined with some of the predictors from the other correlation groups, they provide a better explanation of the species distribution then bathymetry and shore distance do.

Unlike the effect of spatial thinning, using a target-group background resulted in large differences in predictor relevance. As most of the species occurrence records are located along the coast, the target-group background, which is a subsample of it, is expected to have the same bias resulting in a lower relevance of shore distance and bathymetry. These results confirm the importance of background selection on SDMs (Acevedo, Jiménez-Valverde, Lobo, & Real, 2012; Barbet-Massin, Jiguet, Albert, & Thuiller, 2012; Chefaoui & Lobo, 2008; Phillips et al., 2009; Senay, Worner, & Ikeda, 2013; Smith, 2013; VanDerWal et al., 2009). It is therefore recommended to investigate the impact of alternative pseudo-absence selection methods in future studies. Note that in general it is advised to create a species-specific target-group with occurrence records from the same sampling campaign(s) and/or from similar species, reflecting the sampling bias of the species modelled (Phillips et al., 2009).

We explored the impact of several parameter settings on predictor selection; however, the potential analyses are by no means exhaustive. For example, the regularization parameter and the complexity of the features in MaxEnt, the number of trees fitted in random forests and the usage of polynomial features in GLM were kept constant or were not explored. It is likely that applying species-specific tuning of the algorithms will not only impact model performance but also affect the predictor selection (Anderson & Gonzalez, 2011; Merow et al., 2014). We also expect that the objective and scale of the study will impact the relevance of predictors (Nyström Sandman et al., 2013; Pittman & Brown, 2011).

Similar to Barbet-Massin and Jetz (2014), predictor relevance in this study was assessed based on ranking the model performance of all combinations of predictor on the evaluation dataset. This approach differs from assessing predictor importance, which is a measure of the relative contribution of a variable within a model. Predictor importance within a model is commonly assessed, for example, in MaxEnt and BIOMOD2, by randomly permuting the values of the different predictors and measuring the drop in model performance. A further study is needed to uncover the relationship between these two metrics. Using the predictor importance within a model for a subset of all combinations might provide sufficient information for estimating the predictor relevance and thus significantly reduce the number of models that have to be built.

From a species perspective, we noted that the taxonomy and the traits of a species have an influence on the relevance of predictors. The overarching pattern of predictor relevance holds up across traits, but some marked differences in predictor relevance were found for shore distance and bathymetry and to a lesser extent for diffuse attenuation, phosphate, nitrate and silicate. To some extent, these differences are intuitive. For example, subdividing the taxa between oceanic and neritic species results in a higher relevance of shore distance for neritic species. Likewise, SST range is less relevant for tropical and polar species, because low and high latitudes typically exhibit very little annual sea surface temperature fluctuations compared to mid-latitudes. Despite some pronounced differences across traits, trends for inorganic nutrients (nitrate, phosphate, silicate) are less easily explained.

4.2 | Benchmark dataset

Inspired by the widespread use of benchmark datasets in machine learning and other computational fields, we set out to create MarineSPEED. Although a series of papers was published using the same set of 226 terrestrial species (e.g., Elith et al., 2006; Guisan et al., 2007; Hijmans, 2012; Phillips et al., 2009), most studies discussing new methods related to SDM use a small set of different species. Moreover, while the resulting algorithm and methods are regularly made available through ready to use R packages or desktop programs, the species distribution records used in these studies often are not. With the release of MarineSPEED and its associated R package, researchers can download all occurrences, background records and cross-validation datasets.

The marine character of the dataset is ideally suited for the study of methodological issues and parameterizations for distribution modelling of non-terrestrial species. This is necessary as the marine environment poses its own challenges for SDM (Bentlage et al., 2013; Dambach & Rödder, 2011; Kaschner, Watson, Trites, & Pauly, 2006; MacLeod et al., 2008; Robinson et al., 2011). Species distribution records from public databases contain a combination of opportunistic records and systematic sampling campaigns. They show large biases in amount and location of occurrences where the coastal areas are often more intensely sampled than offshore areas. The lower detectability of marine species in combination with the wide extent of the marine environment leads to false absences and a general lack of distribution records in comparison with the real world range extent of marine species. MacLeod et al. (2008) found that in contrast to the terrestrial environment, presence–absence methods do not perform better than presence-only methods in the marine environment. Although absences are rarely reported for marine species and not included in MarineSPEED, this study could be confirmed using estimated absence data for species included in systematic surveys in OBIS (Coro et al., 2016).

4.3 | Applications

Combining the MARINESPEED R package with one of the numerous SDM packages like BIOMOD2, DISMO, SDM or ZOON, other machine learning packages like CARET, GBM, RANDOMFOREST or XGBOOST and the general R ecosystem allows for numerous applications.

While several papers have compared the performance of SDM algorithms (e.g., Elith et al., 2006; Liu, White, & Newell, 2011; Lorena et al., 2011; Meynard & Quinn, 2007; Tsoar, Allouche, Steinitz, Rotem, & Kadmon, 2007), new SDM modelling algorithms are regularly released (e.g., MaxLike (Royle, Chandler, Yackulic, & Nichols, 2012), Plateau (Brewer, O'Hara, Anderson, & Ohlemüller, 2016), GRaF (Golding & Purse, 2016)). Consistent usage of MarineSPEED to explore the performance of modelling algorithms would allow for a direct comparison of the strengths and weaknesses of them. On top of this, SDM algorithms benefit from species-specific parameter settings (Anderson & Gonzalez, 2011; Merow, Smith, & Silander, 2013; Shcheglovitova & Anderson, 2013), but useful ranges for the different parameters are unknown for these newer modelling algorithms.

Over the years, numerous studies have been published on methods for correcting sample selection bias (e.g., Aiello-Lammens et al., 2015;

Barnes et al., 2014; Boria, Olson, Goodman, & Anderson, 2014; Dudík, Schapire, & Phillips, 2005; Fernández & Nakamura, 2015; Phillips et al., 2009; Ranc et al., 2016; Varela, Anderson, García-Valdés, & Fernández-González, 2014) and selecting pseudo-absence records (e.g., Acevedo et al., 2012; Assis et al., 2015; Barbet-Massin et al., 2012; Lobo & Tognelli, 2011; Senay et al., 2013; Wisz & Guisan, 2009). Comparing these techniques with MarineSPEED can result in guidelines for sampling bias mitigation and pseudo-absence selection in the marine environment.

Next to the availability of marine species with environmental data and traits we expect that the MARINESPEED R package, with its implementation of cross-validation methods, to be a useful tool for SDM. Installation instructions, data downloads and species information can be found at <http://marinespeed.org/>.

ACKNOWLEDGEMENTS

The research was carried out with financial support from the ERANET INVASIVES project (EU FP7 SEAS-ERA/INVASIVES SD/ER/010) and financial, data & infrastructure support provided by VLIZ as part of the Flemish contribution to the LifeWatch ESFRI. The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, the Hercules Foundation and the Flemish Government—department EWI. We thank An Verfaillie, Taban Mestdag and Sara Martinez for contributing to the creation of the dataset. We further thank Bart Vanhoorne and Leen Vandepitte from WoRMS and Pieter Provoost from OBIS for providing support.

DATA ACCESSIBILITY

The benchmark data can be downloaded from http://marinespeed.org/. The release version of the R package is on CRAN, and the latest development version can be found at <https://github.com/lifewatch/marinespeed>.

ORCID

Samuel Bosch http://orcid.org/0000-0002-2514-0283
Olivier De Clerck http://orcid.org/0000-0002-3699-8402

REFERENCES

Acevedo, P., Jiménez-Valverde, A., Lobo, J. M., & Real, R. (2012). Delimiting the geographical background in species distribution modelling. *Journal of Biogeography*, *39*, 1383–1390.

Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B., & Anderson, R. P. (2015). spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, *38*, 541–545.

Anderson, R. P., & Gonzalez, I. (2011). Species-specific tuning increases robustness to sampling bias in models of species distributions: An implementation with Maxent. *Ecological Modelling*, *222*, 2796–2811.

Araújo, M. B., & Guisan, A. (2006). Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, *33*, 1677–1688.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79.

Assis, J., Zupan, M., Nicastro, K. R., Zardi, G. I., McQuaid, C. D., & Serrão, E. A. (2015). Oceanographic conditions limit the spread of a marine invader along Southern African Shores. *PLoS ONE*, *10*, e0128124

Austin, M. P. (2002). Spatial prediction of species distribution: An interface between ecological theory and statistical modelling. *Ecological Modelling*, *157*, 101–118.

Bahn, V., & McGill, B. J. (2007). Can niche-based distribution models outperform spatial interpolation? *Global Ecology and Biogeography*, *16*, 733–742.

Barbet-Massin, M., & Jetz, W. (2014). A 40-year, continent-wide, multispecies assessment of relevant climate predictors for species distribution modelling. *Diversity and Distributions*, *20*, 1285–1295.

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in Ecology and Evolution*, *3*, 327–338.

Barnes, M. A., Jerde, C. L., Wittmann, M. E., Chadderton, W. L., Ding, J., Zhang, J., … Lodge, D. M. (2014). Geographic selection bias of occurrence data influences transferability of invasive Hydrilla verticillata distribution models. *Ecology and Evolution*, *4*, 2584–2593.

Barry, S., & Elith, J. (2006). Error and uncertainty in habitat models. *Journal of Applied Ecology*, *43*, 413–423.

Beaugrand, G., Lenoir, S., Ibañez, F., & Manté, C. (2011). A new model to assess the probability of occurrence of a species, based on presence-only data. *Marine Ecology Progress Series*, *424*, 175–190.

Belanger, C.L., Jablonski, D., Roy, K., Berke, S.K., Krug, A.Z., & Valentine, J.W. (2012). Global environmental predictors of benthic marine biogeographic structure. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 14046–14051.

Bentlage, B., Peterson, A. T., Barve, N., & Cartwright, P. (2013). Plumbing the depths: Extending ecological niche modelling and species distribution modelling in three dimensions. *Global Ecology and Biogeography*, *22*, 952–961.

Booth, T. H., Nix, H. A., Busby, J. R., & Hutchinson, M. F. (2014). BIOCLIM: The first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. *Diversity and Distributions*, *20*, 1–9.

Boria, R. A., Olson, L. E., Goodman, S. M., & Anderson, R. P. (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, *275*, 73–77.

Bosch, S., Tyberghein, L., & De Clerck, O. (2016). *sdmpredictors: Species distribution modelling predictor datasets*. R package version 0.9. Available at: https://github.com/lifewatch/sdmpredictors.

Bradie, J., & Leung, B. (2016). A quantitative synthesis of the importance of variables used in MaxEnt species distribution models. *Journal of Biogeography*, *44*, 1344–1361.

Braunisch, V., Coppes, J., Arlettaz, R., Suchant, R., Schmid, H., & Bollmann, K. (2013). Selecting from correlated climate variables: A major source of uncertainty for predicting species distributions under climate change. *Ecography*, *36*, 971–983.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Brewer, M. J., O'Hara, R. B., Anderson, B. J., & Ohlemüller, R. (2016). Plateau: A new method for ecologically plausible climate envelopes for species distribution modelling. *Methods in Ecology and Evolution*, *7*, 1489–1502.

Bucklin, D. N., Basille, M., Benscoter, A. M., Brandt, L. A., Mazzotti, F. J., Romañach, S. S., … Watling, J. I. (2015). Comparing species distribution models constructed with different subsets of environmental predictors. *Diversity and Distributions*, *21*, 23–35.

Chamberlain, S., Boettiger, C., Karthik, R., Barve, V., & Mcglinn, D. (2016). *rgbif: Interface to the Global Biodiversity Information Facility API*. R package version 0.9.3. Available at: https://github.com/ropensci/rgbif.

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2016). *shiny: Web Application Framework for R*. R package version 0.14.1. Available at: http://shiny.rstudio.com.

Chefaoui, R. M., & Lobo, J. M. (2008). Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecological Modelling*, *210*, 478–486.

Cheung, W. W. L., Lam, V. W. Y., & Pauly, D. (2008). Dynamic bioclimate envelope model to predict climate-induced changes in distribution of marine fishes and invertebrates. *Fisheries Centre Research Report*, *16*(16), 5–50.

Coro, G., Magliozzi, C., Vanden, Berghe. E., Bailly, N., Ellenbroek, A., & Pagano, P. (2016). Estimating absence locations of marine species from data of scientific surveys in OBIS. *Ecological Modelling*, *323*, 61–76.

Cucuringu, M. (2016). Sync-Rank: Robust ranking, constrained ranking and rank aggregation via eigenvector and SDP synchronization. *IEEE Transactions on Network Science and Engineering*, *3*, 58–79.

Dambach, J., & Rödder, D. (2011). Applications and future challenges in marine species distribution modeling. *Aquatic Conservation: Marine and Freshwater Ecosystems*, *21*, 92–100.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., … Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*, 027–046.

Downie, A.-L., von Numers, M., & Boström, C. (2013). Influence of model selection on the predicted distribution of the seagrass Zostera marina. *Estuarine, Coastal and Shelf Science*, *121–122*, 8–19.

Dudík, M., Schapire, R. E., & Phillips, S. J. (2005). Correcting sample selection bias in maximum entropy density estimation. *Advances in Neural Information Processing Systems*, *18*, 323–330.

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., … Zimmermann, N. E. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*, 129–151.

Elith, J., Kearney, M., & Phillips, S. (2010). The art of modelling range-shifting species. *Methods in Ecology and Evolution*, *1*, 330–342.

Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, *40*, 677–697.

Fernández, D., & Nakamura, M. (2015). Estimation of spatial sampling effort based on presence-only data and accessibility. *Ecological Modelling*, *299*, 147–155.

Fitzpatrick, M. C., & Hargrove, W. W. (2009). The projection of species distribution models and the problem of non-analog climate. *Biodiversity and Conservation*, *18*, 2255–2261.

Fogel, F., D'Aspremont, A., & Vojnovic, M. (2016). Spectral ranking using seriation. *Journal of Machine Learning Research*, *17*, 1–45.

Franklin, J. (2009). *Mapping species distributions. Spatial inference and prediction*. Cambridge, USA: Cambridge University Press.

R. Froese & D. Pauly (Eds.). (2017). FishBase. Available from http://www.fishbase.org. Accessed 2017-05-11.

Golding, N., & Purse, B. V. (2016). Fast and flexible Bayesian species distribution modelling using Gaussian processes. *Methods in Ecology and Evolution*, *7*, 598–608.

Guisan, A., Graham, C. H., Elith, J., Huettmann, F., Dudik, M., Ferrier, S., … Zimmermann, N. E. (2007). Sensitivity of predictive species distribution models to change in grain size. *Diversity and Distributions*, *13*, 332–340.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29–36.

Herbrich, R., Minka, T., & Graepel, T. (2006). TrueSkill: A Bayesian skill rating system. *Advances in Neural Information Processing Systems*, *20*, 569–576.

Hijmans, R. J. (2012). Cross-validation of species distribution models: Removing spatial sorting bias and calibration with a null model. *Ecology*, *93*, 679–688.

Hijmans, R. J., Phillips, S., Leathwick, J., & Elith, J. (2016). *dismo: Species Distribution Modeling*. R package version 1.1-1. Available at: http://cran.r-project.org/package=dismo.

Huang, Z., Brooke, B., & Li, J. (2011). Performance of predictive models in marine benthic environments based on predictions of sponge distribution on the Australian continental shelf. *Ecological Informatics*, *6*, 205–216.

Kaschner, K., Watson, R., Trites, A., & Pauly, D. (2006). Mapping world-wide distributions of marine mammal species using a relative environmental suitability (RES) model. *Marine Ecology Progress Series*, *316*, 285–310.

Kearney, M., & Porter, W. (2009). Mechanistic niche modelling: Combining physiological and spatial data to predict species' ranges. *Ecology Letters*, *12*, 334–350.

Kramer-Schadt, S., Niedballa, J., Pilgrim, J. D., Schröder, B., Lindenborn, J., Reinfelder, V., … Wilting, A. (2013). The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions*, *19*, 1366–1379.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*, 18–22.

Liu, C., White, M., & Newell, G. (2011). Measuring and comparing the accuracy of species distribution models with presence-absence data. *Ecography*, *34*, 232–243.

Lobo, J. M., Jiménez-Valverde, A., & Hortal, J. (2010). The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, *33*, 103–114.

Lobo, J. M., & Tognelli, M. F. (2011). Exploring the effects of quantity and location of pseudo-absences and sampling biases on the performance of distribution models with limited point occurrence data. *Journal for Nature Conservation*, *19*, 1–7.

Lorena, A. C., Jacintho, L. F. O., Siqueira, M. F., Giovanni, R., De, Lohmann. L. G., de Carvalho, A. C. P. L. F., & Yamamoto, M. (2011). Comparing machine learning classifiers in potential distribution modelling. *Expert Systems with Applications*, *38*, 5268–5275.

MacLeod, C. D., Mandleberg, L., Schweder, C., Bannon, S. M., & Pierce, G. J. (2008). A comparison of approaches for modelling the occurrence of marine animals. *Hydrobiologia*, *612*, 21–32.

Merckx, B., Steyaert, M., Vanreusel, A., Vincx, M., & Vanaverbeke, J. (2011). Null models reveal preferential sampling, spatial autocorrelation and overfitting in habitat suitability modelling. *Ecological Modelling*, *222*, 588–597.

Merow, C., Smith, M. J., Edwards, T. C., Guisan, A., McMahon, S. M., Normand, S., … Elith, J. (2014). What do we gain from simplicity versus complexity in species distribution models? *Ecography*, *37*, 1267–1281.

Merow, C., Smith, M. J., & Silander, J. A. (2013). A practical guide to MaxEnt for modeling species 'distributions: What it does, and why inputs and settings matter. *Ecography*, *36*, 1058–1069.

Meynard, C. N., & Quinn, J. F. (2007). Predicting species distributions: A critical comparison of the most common statistical models using artificial species. *Journal of Biogeography*, *34*, 1455–1469.

Negahban, S., Oh, S., & Shah, D. (2017). Rank Centrality: Ranking from pairwise comparisons. *Operations Research*, *65*, 266–287.

Nyström Sandman, A., Wikström, S. A., Blomqvist, M., Kautsky, H., & Isaeus, M. (2013). Scale-dependent influence of environmental variables on species distribution: A case study on five coastal benthic species in the Baltic Sea. *Ecography*, *36*, 354–363.

Pacifici, K., Reich, B. J., Miller, D. A. W., Gardner, B., Stauffer, G., Singh, S., … Collazo, J. A. (2017). Integrating multiple data sources in species distribution modeling: A framework for data fusion*. *Ecology*, *98*, 840–850.

Palialexis, A., Georgakarakos, S., Karakassis, I., Lika, K., & Valavanis, V. D. (2011). Prediction of marine species distribution from presence–absence acoustic data: Comparing the fitting efficiency and the predictive capacity of conventional and novel distribution models. *Hydrobiologia*, *670*, 241–266.

M. L. D. Palomares & D. Pauly (Eds). (2017). SeaLifeBase. Available from http://sealifebase.org. Accessed 2017-05-11.

Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M., & Araújo, M. B. (2011). *Ecological niches and geographic distributions (MPB-49)*. Princeton, USA: Princeton University Press.

Petitpierre, B., Broennimann, O., Kueffer, C., Daehler, C., & Guisan, A. (2017). Selecting predictors to maximize the transferability of species distribution models: Lessons from cross-continental plant invasions. *Global Ecology and Biogeography*, *26*, 275–287.

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample selection bias and presence-only

distribution models: Implications for background and pseudo-absence data. *Ecological Applications*, *19*, 181–197.
Phillips, S. J., Dudík, M., & Schapire, R. E. (2004). A maximum entropy approach to species distribution modeling. Twenty-first international conference on Machine learning—ICML '04, 655–662.
Pittman, S. J., & Brown, K. A. (2011). Multi-scale approach for predicting fish species distributions across coral reef seascapes. *PLoS ONE*, *6*, e20583.
Pörtner, H.O. (2002). Climate Variations and the Physiological Basis of Temperature Dependent Biogeography: Systemic to Molecular Hierarchy of Thermal Tolerance in Animals. Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology, 132, 739-761. http://dx.doi.org/10.1016/S1095-6433(02)00045-4
Provoost, P., Bosch, S., & Appeltans, W. (2016). *robis: R client for the OBIS API*. R package version 0.1.5. Available at: https://github.com/iobis/robis.
R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Core Team. Available at: http://www.r-project.org/.
Ranc, N., Santini, L., Rondinini, C., Boitani, L., Poitevin, F., Angerbjörn, A., & Maiorano, L. (2016). Performance tradeoffs in target-group bias correction for species distribution models. *Ecography*, *40*, 1076–1087.
Randin, C. F., Dirnböck, T., Dullinger, S., Zimmermann, N. E., Zappa, M., & Guisan, A. (2006). Are niche-based species distribution models transferable in space? *Journal of Biogeography*, *33*, 1689–1703.
Ready, J., Kaschner, K., South, A. B., Eastwood, P. D., Rees, T., Rius, J., … Froese, R. (2010). Predicting the distributions of marine organisms at the global scale. *Ecological Modelling*, *221*, 467–478.
Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., … Dormann, C. F. (2016). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*, 913–929.
Robinson, L. M., Elith, J., Hobday, A. J., Pearson, R. G., Kendall, B. E., Possingham, H. P., & Richardson, A. J. (2011). Pushing the limits in marine species distribution modelling: Lessons from the land present challenges and opportunities. *Global Ecology and Biogeography*, *20*, 789–802.
Royle, J. A., Chandler, R. B., Yackulic, C., & Nichols, J. D. (2012). Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods in Ecology and Evolution*, *3*, 545–554.
Sbrocco, E. J., & Barber, P. H. (2013). MARSPEC: Ocean climate layers for marine spatial ecology. *Ecology*, *94*, 979.
Senay, S. D., Worner, S. P., & Ikeda, T. (2013). Novel three-step pseudo-absence selection technique for improved species distribution modelling. *PLoS ONE*, *8*, e71218.
Shcheglovitova, M., & Anderson, R. P. (2013). Estimating optimal complexity for ecological niche models: A jackknife approach for species with small sample sizes. *Ecological Modelling*, *269*, 9–17.
Šiaulys, A., & Bučas, M. (2012). Species distribution modelling of benthic invertebrates in the south-eastern Baltic Sea. *Baltica*, *25*, 163–170.
Smith, A. B. (2013). On evaluating species distribution models with random background sites in place of absences when test presences disproportionately sample suitable habitat. *Diversity and Distributions*, *19*, 867–872.
Spalding, M.D., Fox, H.E., Allen, G.R., Davidson, N., Ferdaña, Z. a., Finlayson, M., Halpern, B.S., Jorge, M. a., Lombana, A., Lourie, S. a., Martin, K.D., Mcmanus, E., Molnar, J., Recchia, C. a., & Robertson, J. (2007). Marine Ecoregions of the World: A Bioregionalization of Coastal and Shelf Areas. *BioScience*, *57*, 573–583.
Stirling, D. A., Boulcott, P., Scott, B. E., & Wright, P. J. (2016). Using verified species distribution models to inform the conservation of a rare marine species. *Diversity and Distributions*, *22*, 808–822.
Syfert, M. M., Smith, M. J., & Coomes, D. A. (2013). The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. *PLoS ONE*, *8*, e55158.
Synes, N. W., & Osborne, P. E. (2011). Choice of predictor variables as a source of uncertainty in continental-scale species distribution modelling under climate change. *Global Ecology and Biogeography*, *20*, 904–914.
Tittensor, D. P., Mora, C., Jetz, W., Lotze, H. K., Ricard, D., Berghe, E. V., & Worm, B. (2010). Global patterns and predictors of marine biodiversity across taxa. *Nature*, *466*, 1098–1101.
Torres, L., Read, A., & Halpin, P. (2008). Fine-scale habitat modeling of a top marine predator: Do prey data improve predictive capacity. *Ecological Applications*, *18*, 1702–1717.
Tsoar, A., Allouche, O., Steinitz, O., Rotem, D., & Kadmon, R. (2007). A comparative evaluation of presence-only methods for modelling species distribution. *Diversity and Distributions*, *13*, 397–405.
Tyberghein, L., Verbruggen, H., Pauly, K., Troupin, C., Mineur, F., & De Clerck, O. (2012). Bio-ORACLE: A global environmental dataset for marine species distribution modelling. *Global Ecology and Biogeography*, *21*, 272–281.
VanDerWal, J., Shoo, L. P., Graham, C., & Williams, S. E. (2009). Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? *Ecological Modelling*, *220*, 589–594.
Varela, S., Anderson, R. P., García-Valdés, R., & Fernández-González, F. (2014). Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, *37*, 1084–1091.
Verbruggen, H., Tyberghein, L., Belton, G. S., Mineur, F., Jueterbock, A., Hoarau, G., … De Clerck, O. (2013). Improving transferability of introduced species' distribution models: New tools to forecast the spread of a highly invasive seaweed. *PLoS ONE*, *8*, e68337.
Walther, G.-R., Roques, A., Hulme, P. E., Sykes, M. T., Pysek, P., Kühn, I., … Settele, J. (2009). Alien species in a warmer world: Risks and opportunities. *Trends in Ecology & Evolution*, *24*, 686–693.
Wisz, M. S., & Guisan, A. (2009). Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. *BMC Ecology*, *9*.
WoRMS Editorial Board. (2016). World Register of Marine Species. Available from http://www.marinespecies.org at VLIZ. Accessed 2016-12-20.
Zheng, B., & Agresti, A. (2000). Summarizing the predictive power of a generalized linear model. *Statistics in Medicine*, *19*, 1771–1781.

BIOSKETCH

Samuel Bosch (SB) carries out research on marine species distributions, ranging from quality control in public databases to modelling invasive seaweeds. He feels most comfortable when combining spatial data with programming and ecological modelling.

Author contributions: SB, LT and ODC conceived the ideas; SB and ODC collected the data; SB ran the models and analysed the data. SB and ODC led the writing of the manuscript, to which all authors contributed.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Bosch S, Tyberghein L, Deneudt K, Hernandez F, De Clerck O. In search of relevant predictors for marine species distribution modelling using the MarineSPEED benchmark dataset. *Divers Distrib*. 2017;00:1–14. https://doi.org/10.1111/ddi.12668